Sequence tag–based analysis of microbial population dynamics

Published in final edited form as:
Nat Methods. 2015 March ; 12(3): 223–226. doi:10.1038/nmeth.3253.

STAMP: Sequence tag-based analysis of microbial population dynamics

Sören Abel[1,2], **Pia Abel zur Wiesch**[3,4], **Hsiao-Han Chang**[5], **Brigid M. Davis**[1,2], **Marc Lipsitch**[5], and **Matthew K. Waldor**[1,2,6]

[1]Department of Microbiology and Immunobiology, Harvard Medical School, Boston, Massachusetts, United States of America

[2]Division of Infectious Diseases, Brigham & Women's Hospital, Boston, Massachusetts, United States of America

[3]Department of Global Health and Social Medicine, Harvard Medical School, Boston, Massachusetts, United States of America

[4]Division of Global Health Equity, Brigham and Women's Hospital, Boston, Massachusetts, United States of America

[5]Center for Communicable Disease Dynamics, Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts, United States of America

[6]Howard Hughes Medical Institute, Boston, Massachusetts, United States of America

Abstract

We describe a new method (STAMP) for characterization of pathogen population dynamics during infection. STAMP analyzes the frequency changes of genetically "barcoded" organisms to quantify population bottlenecks and infer the founding population size. Analyses of intra-intestinal *Vibrio cholerae* revealed infection-stage and region-specific host barriers to infection, and unexpectedly showed *V. cholerae* migration counter to intestinal flow. STAMP provides a robust, widely applicable analytical framework for high confidence characterization of *in vivo* microbial dissemination.

A pathogen's population dynamics within a host organism reflect a plethora of factors, including the availability of hospitable niches for colonization, the extent of host barriers to infection, and the pathogen's capacity for replication[1–3]. However, it can be difficult to parse the relative impacts of these factors using traditional approaches, such as enumeration of colony-forming units (cfu) at different times and sites of infection, and such analyses typically require use of a high number of experimental animals. Robust mathematical

Correspondence should be addressed to: M.K.W. (mwaldor@research.bwh.harvard.edu).

AUTHOR CONTRIBUTIONS

S.A., P.A.z.W. and M.K.W. designed experiments. S.A. performed experiments. S.A., P.A.z.W. and H-H. C. analyzed the data. S.A., P.A.z.W., B.D., M.L. and M.K.W. wrote the manuscript. All authors discussed the results and commented on the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

frameworks have been developed to identify and classify events that shape population structures over time based on natural variation in the genetic composition of populations, but these have generally been applied in studies of eukaryotic evolutionary biology in which numerous distinguishable alleles are present[4–7]. The inocula of infectious microbes used in laboratory analyses usually lack sufficient distinguishable alleles for high resolution analysis of pathogen population dynamics. Furthermore, the effects of natural polymorphisms are not necessarily neutral, so it can be difficult to distinguish genetic drift from selection. Artificial tags have been used to create distinguishable pathogens that are more easily analyzed and have equivalent fitness[8–14]. Most recently, sequence "barcodes" have been used as tags in a method termed WITS (wild-type isogenic tagged strains)[12–14]. However, these studies have so far been limited by the use of small numbers of tags, which restrict their resolving power, by the need for specialized mathematical models that require assumptions about the spatiotemporal spread of the pathogen within the host, and by lack of a systematic approach for analysis of tag frequencies in different populations. These limitations are not critical when the size of the founding populations, i.e., bacteria that survive host defenses and subsequently replicate, is very small, e.g., when only one or a few organisms overcome the host defenses and colonize specific tissues or organs. However, they severely constrain the information that can be obtained from more complex founding populations. For example, one very recent study does provide an analysis framework for use with WITS data, based on a stochastic model of tag loss; however, this approach only yields high confidence results when the compartment of interest is seeded by a relatively small (maximum ~10^2) number of organisms[14].

In our work, we have combined classical population analysis frameworks with the power of high-throughput DNA sequencing technology and large libraries of neutrally tagged pathogens to generate a new approach for dissection of microbial population dynamics during infection (STAMP; Sequence Tag-based Analysis of Microbial Populations) that is applicable to analyses of all populations, regardless of their complexity. From the relative abundance (rather than simply presence or absence) of hundreds of individually tagged but otherwise isogenic strains within the infection inoculum and at various times and sites during infection, we can estimate the number of bacteria from the inoculum whose descendants are represented in a population at the time and site of sampling. This number, which we term founding or bottleneck population size (N_b), reflects the stringency of host barriers encountered during infection, and allows the magnitude of such restrictions to be assessed retrospectively, without knowledge of their timing or location. STAMP permits unprecedented high-resolution determination of N_b over an extremely large dynamic range, limited in practice only by the depth of available high throughput sequencing results. We demonstrate STAMP's utility through analysis of the population dynamics of the cholera pathogen, *V. cholerae*, in the infant rabbit model of infection[15].

We hypothesized that N_b could be estimated using approaches for determination of effective population size (N_e)[16], a key parameter for modeling population dynamics, and first confirmed this theory *in silico*, using simulations in which we varied bottleneck size and the number of tags (Supplementary Fig. 1 and Online Methods). These simulations also revealed that 500 tags is sufficient for high confidence determination of N_b (Supplementary

Fig. 1), while 50 tags (more than used in previous analyses) is predicted to yield less robust results, particularly for high values of N_b. To validate our hypothesis experimentally, a library of *V. cholerae* that were individually barcoded with one of ~500 distinct, short sequence tags inserted into a neutral locus on the chromosome was generated (Fig. 1a and Supplementary Fig. 2). We sampled defined numbers of bacteria (10^1–10^7 cfu) to simulate bottleneck events *in vitro*, and assessed whether changes in the frequency of individual tags relative to the initial library could be used to estimate N_b after the sample was expanded on agar plates (Fig. 1a,b). Using several approaches for sequencing-based estimation of N_b, we identified parameters for optimal analysis (Supplementary Fig. 3). The most extensive correlation between sequence-based estimation of N_b and associated experimentally determined bacterial load were found using equations from Krimbas & Tsakas (Fig. 1b and Supplementary Fig. 3)[5]:

$$\hat{F}=\frac{1}{k}\sum_{i=1}^{k}\frac{(f_{i,s}-f_{i,0})^2}{f_{i,0}(1-f_{i,0})} \quad (1)$$

and

$$N_b \approx N_e=\frac{g}{\hat{F}-\frac{1}{S_0}-\frac{1}{S_s}} \quad (2)$$

where k is the total number of distinct alleles (i.e., number of unique tags), $f_{i,0}$, the frequency of allele i at time 0, $f_{i,s}$, the frequency of allele i at sampling, g, the number of generations during competitive growth, S_0 and S_s the sample size used to determine the population composition (i.e., the number of sequence reads) at time 0 or at sampling, respectively (Online Methods). There was a very high correlation ($R^2 = 0.99$) between the estimated N_b and the associated bacterial load over a range of ~5 orders of magnitude (Fig. 1b), suggesting that this approach could enable accurate assessment of population bottlenecks over this range *in vivo*. Populations with N_b of ~10^6 and higher (corresponding to very "wide" bottlenecks) were indistinguishable from each other in our analyses because we obtained fewer than 10^6 sequence reads per sample therefore sequencing depth itself act as a bottleneck and limits the resolution. For smaller populations, our calculations yielded an estimated N_b slightly lower than the cfu-based value. Consequently, this *in vitro* data was used as a calibration curve for our subsequent *in vivo* experiments and the corrected *in vivo* values are denoted N_b'.

Like infected humans, infant rabbits orogastrically infected with *V. cholerae* develop severe and potentially fatal diarrhea, due to the pathogen's colonization of the small intestine (SI) and subsequent secretion of cholera toxin[15]. We harvested bacteria from intestinal homogenates of animals infected with 10^9 cfu of our tagged library at 20 h post-infection (PI), at which point the animals exhibit severe cholera-like diarrhea, and found that the estimated N_b' was ≤ 10^5 for all sections of the gastrointestinal (GI) tract (Fig. 1c). This indicates that only a small subset of the inoculum establishes infection and that the original population size (10^9) is reduced by ~4 orders of magnitude. However, the *V. cholerae* that successfully found the population replicate robustly, such that by 20 h PI the absolute

bacterial load (recoverable cfu) is markedly higher than the associated N_b' value for all sites (Fig. 1c). Thus, bacterial replication masks the earlier effect of host bottlenecks on recoverable cfu; however, the prior effects of bottlenecks can be detected through tag-based estimation of N_b. The N_b' value derived from STAMP is similar to the number of unique transposon insertion mutants recovered after inoculation with a complex transposon library, which provides an alternative estimate of the lower limit for N_b (Supplementary Fig. 4) and estimates from the literature[17]. In addition to rabbits, we also determined the founding population size in infant mice. N_b' was slightly lower (~3.9×10^4) in mice than in rabbits (Supplementary Fig. 5), which is consistent with bottleneck determinations by direct observation[18] and using complex transposon libraries[19].

Surprisingly, the founding population size was not uniform across the GI tract. In the mid SI (P8–10 and I2) at 20 h PI, N_b' was nearly 1000-fold lower than in the proximal (P1–7) and distal SI (I3) (Fig. 1c,d and Supplementary Fig. 6). These distinct N_b' values reveal that the sites where *V. cholerae* can establish infection are not uniformly distributed among SI sections, an insight that is not evident from enumeration of bacterial loads in each section at this time point (Fig. 1c). The smaller size of the founding population in the mid (P8–10 and I2, ~10^2) vs. the distal (I3, ~10^5) SI indicates that populations along the intestinal tract are not necessarily derived from or continually replenished by populations at “upstream” sites, but instead can maintain distinct identities. It is not yet known why such a small subset of the *V. cholerae* inoculum establishes residency in the middle section of the SI, particularly since enumeration of cfu indicates that this region fully supports subsequent bacterial replication.

To gain further insight into the kinetics and directionality of *V. cholerae* spread within the intestinal tract, we estimated N_b' for a variety of intestinal sites (Fig. 1d) at three distinct phases of infection (Fig. 2). During the early phase of infection (~2 h PI), N_b' and bacterial load were relatively high (~10^5 and ~10^7 respectively) in all sections, indicating that the bacteria quickly disseminate from the site of inoculation in the stomach to the distal intestine. It is also evident from the fact that N_b' < cfu at all sites that bacterial replication *in vivo* begins within the first 2 h of infection, rather than requiring a longer period of phenotypic adaptation. In the middle phase of infection (~7 h PI), both N_b' and cfu were reduced (relative to the early phase) for most regions of the intestine, suggesting that most of the inoculum is cleared relatively early after inoculation. Furthermore, cfu are no longer recoverable from the stomach at the middle phase; therefore this site cannot re-seed downstream intestinal segments later in infection. However, during this middle phase, N_b' for the distal SI (I3) remained at ~10^5, suggesting that I3 contains abundant niches that are permissive for *V. cholerae* growth from the onset of infection onwards. Given the absence of viable bacteria in the stomach at ~7 h PI and the low N_b' (~10^2) and bacterial load of *V. cholerae* in upstream sections in this phase, our observation that N_b' for the proximal SI (I1) increases to ~10^4 at the late phase of infection (~20 h PI) likely reflects reseeding of I1 with bacteria from I3. Notably, consumption of contaminated food or stool is not required for reseeding of I1 (Supplementary Fig. 7), strongly suggesting that the increased diversity of founders in I1 late in the infection results (at least in part) from migration of bacteria from the distal to the proximal region of the GI tract, i.e., movement counter to the usual direction

of intestinal flow. Such migration may be aided by the onset of cholera toxin-induced fluid secretion, which typically occurs around this time (Supplementary Fig. 8). While the biological significance of the backward migration of *V. cholerae* is not clear, these observations provide new insight into *V. cholerae*-host dynamics: the change in I1 N_b' indicates either that the number of permissive niches for *V. cholerae* growth in this region changes during the course of infection or that bacteria become better adapted to replicate in I1 after initial growth in I3. Intriguingly, the backward migration of *V. cholerae* from I3 to I1 does not substantially alter the number of founders in I2, which remains at ~10^2 during the late phase of infection. Despite the difference between N_b' of I1 and I2, these segments contained similar numbers of recoverable *V. cholerae* (~10^8 cfu) in the late phase of infection, suggesting that both sites allow for robust replication once bacteria establish a foothold.

Large numbers of tags also enable estimation of the "genetic distance" separating pathogen populations from different intestinal segments and phases of infection by comparative analyses of barcode frequencies based on the chord distance[20]. The results from these analyses were congruent with our conclusions regarding *V. cholerae*'s intra-intestinal dynamics (Fig. 2b). All populations in the SI were closely related in the early phase of infection, probably reflecting relatively even spread of the inoculum, but the genetic distance between them increased by the middle phase. However, by the late phase, the populations in I1 and I3 were again very closely related, supporting our hypothesis that the former is derived from the latter. Thus, our STAMP-based spatial and temporal analyses of *V. cholerae* population dynamics *in vivo* reveal unexpected complexity in pathogen migration patterns and in the host landscape, which largely could not be deduced from traditional approaches to investigation of colonization[18]. Furthermore, the extremely low variability of STAMP-based analyses *in vitro* (Fig. 1b) gives us confidence that the range of N_b values estimated for intestinal sites reflect genuine inter-section and inter-animal variability, rather than technical limitations of the analysis.

In summary, STAMP is a conceptually novel approach that unites molecular biology and next-generation sequencing with equations from classical population genetics to enable quantitative assessment of founding population sizes and retrospective analysis of cell migration patterns, as outlined for *V. cholerae* in figure 2c. By using STAMP and the genetic "relatedness" of the neutrally tagged strains we discovered that i) the bottleneck sizes in the rabbit intestine change during infection - a novel concept in host-pathogen interactions and ii) *V. cholerae* undergoes retrograde movement in the rabbit intestine. None of these insights are foreshadowed in earlier work, nor would their discovery have been possible with previous WITS-based approaches. In contrast to earlier studies[8–14,17–19], STAMP enables systematic and robust analysis of populations with a large number of barcodes, which is critical for high confidence analyses (Supplementary Fig. 2), and it can resolve bottlenecks over a dynamic range orders of magnitude larger than in previous analyses (Fig. 1b). Thus, STAMP enabled measurement of *V. cholerae's* wide bottleneck, which would not have been feasible with fewer tags and previous WITS-based systems unless many more animals were used. Finally, STAMP's power allows us to identify bottleneck sizes with very small technical error based on a single animal as opposed to

animal population averages (Supplementary Fig. 1), thereby allowing us to approximate biological variance between hosts. STAMP's analysis framework should be universally applicable; it does not have to be tailored to specific settings or organisms and does not require prior knowledge of pathogens' migration patterns within a host.

Thus, STAMP will be applicable to investigation of the *in vivo* population dynamics of diverse bacterial pathogens as well as of host, microbiota, and pathogen factors that govern these dynamics. Such analyses may be particularly interesting for pathogens that disseminate through uncharacterized bottlenecks to secondary sites of infection or for quantification of pathogen transmission between hosts. Additionally, the analytical approach underlying STAMP is equally valid for *in vivo* studies of viruses, parasites, or other organisms that can be barcoded or equivalently tagged. Finally, if coupled with high throughput approaches for tagging eukaryotic cells, STAMP's analytical framework could be used to dissect eukaryotic cell population dynamics, e.g., in models of stem cell dissemination, immune cell maturation, or cancer metastasis.

ONLINE METHODS

Media and growth conditions

All *V. cholerae* strains used here were generated in a streptomycin-resistant mutant of *V. cholerae* El Tor O1 Inaba strain C6706[21]. Bacteria were grown in LB-medium (Difco) supplemented with antibiotic when necessary at 37°C. Carbenicillin (Sigma Aldrich) was used at a final concentration of 50 μg/ml (LB-Carb), streptomycin (Sigma Aldrich) at 200 μg/ml (LB-Strep). Growth curve analyses were conducted in a Bioscreen C growth plate reader and 100-well honeycomb plates (Oy Growth Curves Ab Ltd.), measuring the absorption at 600 nm in 10 min intervals.

Construction of the barcoded *V. cholerae* library

All cloning procedures were conducted using isothermal assembly[22]. Table S1 contains the sequences of all primers used in this study. The plasmids pSoA160 and pSoA158.mix, used for generating the tagged *V. cholerae* library, were created as follows:

1. pSoA160: A 719 bp fragment from pMK2010[23] containing the *ccdB* toxin was amplified with primers P78 and P79 and inserted into pGP704,a suicide plasmid for *V. cholerae* carrying a beta lactamase gene, at the SacI and XbaI sites, yielding pSoA160. The correct plasmid sequence was confirmed by sequencing.
2. pSoA158.mix: A ~1055 bp fragment of VC0610 that included 93 bp of the intergenic region between VC0610 and VC0611 was amplified using primer P110 that contained a 30 bp stretch of random sequence and P80 and inserted into pSoA160 at the SacI and XbaI sites, yielding pSoA158.mix. The correct plasmid sequence of 24 individual colonies was confirmed by sequencing.

pSoA158.mix was transferred to *V. cholerae* by conjugation with SM10 lambda pir[24]. Transconjugants that successfully integrated the plasmid into the genome by homologous recombination were selected with streptomycin and carbenicillin. After 93 of 93 tested colonies were found to have the correct insertion of pSoA158 by PCR using primers P9 and

P10, the remaining colonies were washed of the plate with LB-Carb and pooled. After addition of 10 % DMSO, the pooled library of tagged *V. cholerae* was aliquoted and stored at −80 °C. The integrated plasmid does not affect bacterial fitness and is stably integrated for at least 20 h without selection (Supplementary Fig. 2).

Animal infections

All animal protocols were reviewed and approved by the Harvard Medical Area Standing Committee on Animals. To prepare the inoculum of sequence tagged *V. cholerae*, a frozen aliquot of the library (1 ml, OD_{600} ~10) was diluted 1:10 in LB-Carb and grown for 3 h with shaking and then harvested by centrifugation (5,000 ×g, room temperature, 10 min) and resuspended in sodium bicarbonate solution (2.5 g in 100 ml; pH 9) containing green food dye (FD&C yellow 5 and FD&C blue 1; McCormick).

For infant rabbits experiments, 2–3 day old male and female New Zealand White infant rabbits (Pine Acre Rabbitry) were treated with Zantac (ranitidine-hydrochloride; GlaxoSmithKline) by intraperitoneal injection (2 μg/g body weight) 3 h prior to infection. 0.5 ml of the inoculum was used to infect animals by gavage using a size 5 French catheter (Arrow International). Unless otherwise stated, infant rabbits were infected with 10^9 cfu *V. cholerae*, euthanized at 20 h PI and housed with their mother and littermates for the duration of the experiment. For time-course experiments, infected animals were grouped into early (~2 h PI), middle (~7 h PI), and late (~20 h PI) phase of the disease based on the sampling time post-infection, disease symptoms, and the extent of *V. cholerae* spread. The late phase is characterized by severe diarrhea and significant accumulation of cecal fluid, the middle phase by very little cecal fluid and absence of bacteria in the stomach, and the early phase by the presence of *V. cholerae* in the stomach and no cecal fluid accumulation.

For suckling mice experiments, 5 day old male and female C57BL/6 mice (Charles River) were separated from their mothers 1 h prior to inoculation with *V. cholerae*. Then, they were intragastrically inoculated with 50 μl of the inoculum using 0.28 mm diameter polyethylene tubing (Becton Dickson). Suckling mice were infected with 10^6 *V. cholerae* cfu and euthanized at 24 h PI. We did not compare groups of animals in our experiments, therefore randomization and blinding the investigator to group allocation was not necessary.

At necropsy, the stomach content was collected and the entire intestinal tract from the duodenum to the rectum was removed. The cecal fluid (Cf; if available) was harvested with a syringe and 26G needle (Becton Dickson) and tissue samples (small intestine (I1, P1–10, I2, I3), cecum tissue (Ce), colon (Co)) were gathered as indicated (Fig. 1d). Tissue samples were homogenized in 1ml sterile phosphate-buffered saline (PBS) using a mini-beadbeater-16 and 3.2 mm stainless steel beads (BioSpec Products Inc.). A total of 750 μl of each sample, including stomach content, cecal fluid and three replicate samples of the inoculum were spread on three separate LB-Carb plates and grown for ~18 h to be harvested for N_b estimation. Bacterial load (cfu) was enumerated by plating serial dilutions. A graphical overview of the experimental setup is depicted in figure 1a.

DNA sample preparation and sequencing

To sequence the barcodes, bacterial colonies were washed off the LB-Carb plates with cold PBS and triplicate samples were pooled. Genomic DNA was extracted from samples containing ~3×10^{10} cells using the Wizard Genomic DNA Purification Kit (Promega). A 313 bp fragment containing the tag region was amplified with primer P47 and one of the following primers P48, P51–P73 that contain complementary sequences to Illumina's P5 and P7 grafting primers, respectively, and TruSeq index barcodes using ~2 μg of genomic DNA as template. PCRs were performed in triplicates and the PCR products were pooled before purification with a MinElute PCR Purification Kit (Qiagen).

The DNA was quantified by a Qubit 2.0 fluorometer and Qubit dsDNA HS Assay kit (Life Technologies) and by quantitative PCR with primer P74 and P75 using a Step One Plus Real-Time-PCR machine and Fast SYBR Green Master Mix (Applied Biosystems).

The amplicon libraries were combined in an equimolar fashion and sequenced on an Illumina MiSeq sequencer using a 50 cycle V2 MiSeq reagent kit (Illumina) with custom sequencing primer P49. The libraries were clustered to a density of ~10^6 mm^{-2}. Image analysis, base calling, data quality assessment and de-multiplexing were performed on the MiSeq instrument.

Using reaper-12-340[25] all sequences that contained undefined base calls (N) or did not contain the 14bp of the constant region directly following the random sequence tag were discarded (on average 15.3 % (+/− 9.6 (SD)) of the sequenced reads). In the remaining sequences, the constant region as well as all the following sequence was trimmed. The sequences were converted to FASTA format with convert_fastaqual_fastq.py from QIIME 1.6.0[26] and clustered as well as enumerated with pick_otus.py using uclust[27] and a sequence similarity threshold. The effect of different thresholds for clustering on the N_b estimation was tested; a threshold of 0.9 performed best and was used throughout the study (Supplementary Fig. 3). For each cluster, the most abundant sequence was picked as representative with pick_rep_set.py. In order to remove remaining non-specific tags, all clusters that were not represented in the INOC54 reference set were discarded (see below). The reproducibility of sequencing results was confirmed by comparing the sequences of the same inoculum sample or different independent inocula, re-sequenced on the same or separate sequencing runs (Supplementary Fig. 10). A graphical overview of the analysis setup is depicted in supplementary figure 9.

INOC54 reference set

Sequences from 54 independently sequenced inocula samples were analyzed as described above. However, after the trimming with reaper, all sequences that contained base calls with a quality score below Q30 were discarded. Clusters that were present in 53 out of 54 inocula were included in the INOC54 reference set. A list of all tags is given in table S2.

Bottleneck population size estimation

In an idealized experiment, to answer the question, "How many *V. cholerae* cells from the inoculum have descendants in the population of an intestinal segment sampled at time *t*?",

we would inoculate rabbits with a *V. cholerae* population in which each cell carried a unique tag; harvest the entire population in a segment of intestine at time *t*, and count the number of tags therein. This was not possible for technical and logistical reasons, so we employed mathematical techniques developed in population genetics, specifically the estimation of the effective population size (N_e)[16] based on temporal allele frequency data, to estimate this quantity which we call the founding population size or bottleneck size (N_b). We assume that the changes in allele frequencies are introduced by genetic drift, i.e., by random survival of pathogens that pass through a population bottleneck, however other sources of changes in allele frequency (e.g., niche specific differences in growth rates) can potentially confound the analysis. We hypothesized that applying an estimation formula for N_e would provide a very good estimate of N_b.

While we cannot observe the tag diversity of the *V. cholerae* population at each point in space and time during an infection, we hypothesized that N_b could be estimated by applying a formula to estimate N_e under the simplifying assumption that a single-step bottleneck had occurred, reducing the diversity in the inoculum down to that observed at the sampling time. Methods for estimating N_e also permit correction for the fact that not every member of the population at time *t* is sampled[5,16]. In reality, the loss of diversity from the inoculum probably occurred in every one of several bacterial generations before sampling, though it was likely concentrated in one or a few generations because we observed severe bottlenecks followed by a robust expansion of the pathogen population in the host (Fig. 2). However, because our goal is to determine the number of bacteria that have descendants in the population, N_b, at a given time and location rather than to map each step of the population constriction, we set the generation (*g*) in equation (1) to one to summarize the loss of diversity as having all occurred in a single step. Importantly, the simulation results shown in supplementary figure 1 confirm that the estimates obtained for N_b using our approach accurately reflect the true bottleneck population size. This approach is further supported by the excellent fit of the *in vitro* calibration curve (Fig. 1) and the finding that the N_b estimates remain constant throughout the infection in several sections of the GI-tract (Fig. 2). Several population genetic methods were used to estimate N_b (Supplementary Fig. 3 and data not shown)[4–7]. The best correlation between experimental cfu and estimated N_b was achieved by using equations from Krimbas & Tsakas which was then employed throughout this study (see equation 1 and 2). The R code to estimate N_b is available upon request.

In vitro calibration curve

The inoculum was prepared as described above in triplicates. For each sample, independent 1:10 dilution series in PBS were prepared, spread on LB-Carb plates and grown for ~18 h. The bacterial load was determined and the colonies were harvested for N_b estimation as described above. The data were spline interpolated using the spline function in R [http://www.R-project.org/] with a step width of 0.01 log units and the median as well as 95 % confidence interval was determined. None of the tested N_b estimation methods was perfect, therefore all estimated N_b were corrected using the calibration curve and the resulting values were denoted N_b'. Additionally, when the bottleneck size is larger than the number of obtained sequences, the sampling error becomes larger than the level of genetic drift, which is used for estimating N_b and N_b can no longer be estimated accurately. Therefore a

resolution limit is given in all figures. It depicts the N_b' estimate for which it was still possible to calculate the median and both upper as well as lower confidence boundaries from the *in vitro* calibration data. For given values above the detection limit, the median and lower confidence boundary, but not the upper confidence boundary could be determined.

Statistics

The majority of data was not normally distributed and therefore non-parametric tests were utilized. Wilcoxon signed rank tests (paired data) or rank sum tests (unpaired data) were used to compare two groups and Kruskal-Wallis rank sum test for more than two groups. R was used for all statistical analysis.

Genetic distance

Genetic distances were calculated by the Cavalli-Sforza chord distance method[20] using tag frequency distributions in populations according to:

$$D_{ch}=\frac{2\sqrt{2}}{\pi}\sqrt{1-cos\theta} \quad (3)$$

and

$$cos\theta=\sum_{i=1}^{k}\sqrt{f_{P1,i}f_{P2,i}} \quad (4)$$

where D_{ch} is the chord distance, k is the total number of distinct alleles (number of unique tags), and $f_{P1,i}$ and $f_{P2,i}$ are the frequencies of allele i in population 1 and population 2, respectively. We assume that populations that are in exchange with each other or have only been separated recently are composed of similar relative amounts of organisms carrying individual barcode tags.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank D. Munera for help with animal experiments, M. Chao for help with transposon insertion data, T. Lieberman for discussions and S. Lory, L. Comstock as well as members of the Waldor lab for comments on the manuscript. This work was supported by Swiss Foundation for Grants in Biology and Medicine (www.samw.ch) grant PASMP3_142724/1 (S.A.), Swiss National Science Foundation (www.snf.ch) grant PBEZP3_140163 and German Academic Exchange Service (www.daad.org) grant D/11/45747 (P.A.z.W.), the National Institute of General Medical Sciences of the US National Institutes of Health award number U54GM088558 (H-H.C. and M.L.), US National Institutes of Health grant R37 AI – 042347 and Howard Hughes Medical Institute (M.K.W.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript and the content is solely the responsibility of the authors and does not necessarily represent the official views of the funders.

References

1. Levin BR, Lipsitch M, Bonhoeffer S. Science. 1999; 283:806–809. [PubMed: 9933155]
2. Gutiérrez S, Michalakis Y, Blanc S. Curr Opin Virol. 2012; 2:546–555. [PubMed: 22921636]

3. Watson KG, Holden DW. Cell Microbiol. 2010; 12:1389–1397. [PubMed: 20731667]
4. Wright S. Nature. 1950; 166:247–249. [PubMed: 15439261]
5. Krimbas CB, Tsakas S. Evolution. 1971; 25:454–460.
6. Nei M, Tajima F. Genetics. 1981; 98:625–640. [PubMed: 17249104]
7. Pollak E. Genetics. 1983; 104:531–548. [PubMed: 17246147]
8. Moxon ER, Murphy PA. Proc Natl Acad Sci U S A. 1978; 75:1534–1536. [PubMed: 306628]
9. Margolis E, Levin BR. J Infect Dis. 2007; 196:1068–1075. [PubMed: 17763330]
10. Barnes PD, Bergman MA, Mecsas J, Isberg RR. J Exp Med. 2006; 203:1591–1601. [PubMed: 16754724]
11. Li Y, Thompson CM, Trzciński K, Lipsitch M. Infect Immun. 2013; 81:4534–4543. [PubMed: 24082074]
12. Grant AJ, et al. PLoS Biol. 2008; 6:e74. [PubMed: 18399718]
13. Kaiser P, Slack E, Grant AJ, Hardt WD, Regoes RR. PLoS Pathog. 2013; 9:e1003532. [PubMed: 24068916]
14. Lim CH, et al. PLoS Pathog. 2014; 10:e1004270. [PubMed: 25079958]
15. Ritchie JM, Rui H, Bronson RT, Waldor MK. mBio. 2010; 1:e00047–10. [PubMed: 20689747]
16. Charlesworth B. Nat Rev Genet. 2009; 10:195–205. [PubMed: 19204717]
17. Fu Y, Waldor MK, Mekalanos JJ. Cell Host Microbe. 2013; 14:652–663. [PubMed: 24331463]
18. Angelichio MJ, Spector J, Waldor MK, Camilli A. Infect Immun. 1999; 67:3733–3739. [PubMed: 10417131]
19. Chiang SL, Mekalanos JJ. Mol Microbiol. 1998; 27:797–805. [PubMed: 9515705]
20. Cavalli-Sforza LL, Edwards AW. Am J Hum Genet. 1967; 19:233–257. [PubMed: 6026583]
21. Thelin KH, Taylor RK. Infect Immun. 1996; 64:2853–2856. [PubMed: 8698524]
22. Gibson DG, et al. Nat Methods. 2009; 6:343–345. [PubMed: 19363495]
23. House BL, Mortimer MW, Kahn ML. Appl Environ Microbiol. 2004; 70:2806–2815. [PubMed: 15128536]
24. Simon R, Priefer U, Pühler A. Bio/Technology. 1983; 1:784–791.
25. Davis MPA, van Dongen S, Abreu-Goodger C, Bartonicek N, Enright AJ. Methods San Diego Calif. 2013; 63:41–49.
26. Caporaso JG, et al. Nat Methods. 2010; 7:335–336. [PubMed: 20383131]
27. Edgar RC. Bioinforma Oxf Engl. 2010; 26:2460–2461.
28. Pritchard JR, et al. PLoS Genet. 2014; 10:e1004782. [PubMed: 25375795]

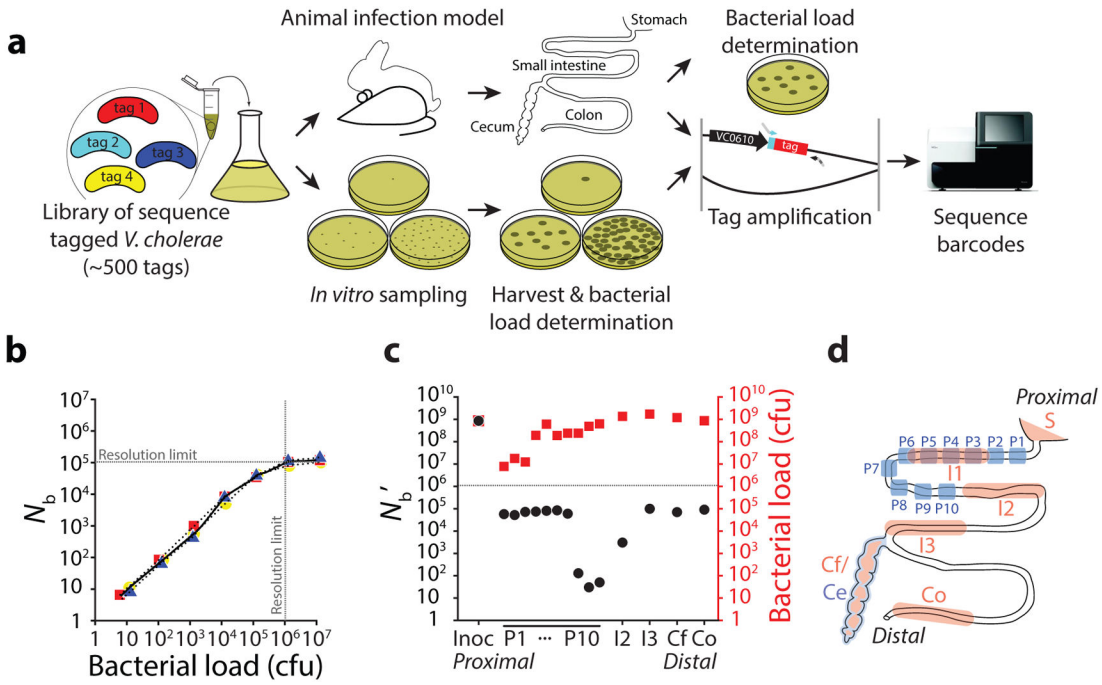

Figure 1.

The barriers to *V. cholerae* infection are heterogeneous along the intestine. (**a**) Schematic overview of the experimental setup. (**b**) *In vitro* calibration curve. Correlation between experimentally determined bottleneck population size (bacterial load) and estimated bottleneck size (N_b) by STAMP. The red, blue and yellow symbols represent biologically independent samples. The solid line indicates the median; the dashed, black lines indicate the 95 % confidence interval. The dashed, grey lines marks the resolution limit for N_b estimation. (**c**) Representative example of bottleneck populations corrected with the calibration curve (N_b', black dots) and bacterial load (cfu, red squares) at 20 h post-infection throughout the gastro-intestinal tract of a single animal after inoculation with 10^9 cfu barcoded *V. cholerae*. An additional example is shown in supplementary figure 6. (**d**) Sampling sites used in this study are indicated in light red or blue; S: stomach content, P1–P10, proximal SI sections used in figure 1, I1: proximal SI section used in figures 2 and 3, I2: middle SI, I3: distal SI, Ce: cecum tissue, Cf: cecal fluid, Co: colon.

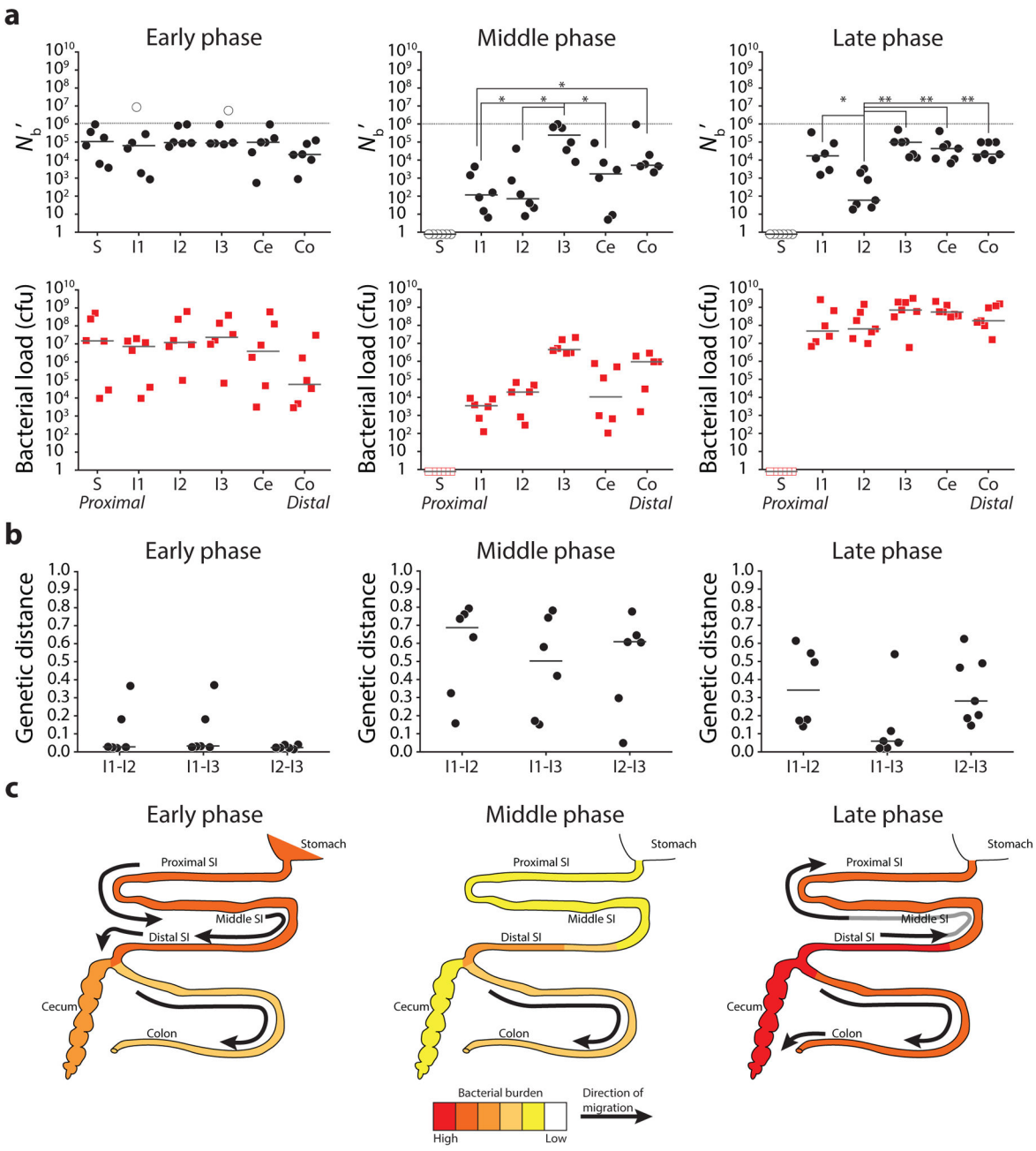

Figure 2.

Spatial and temporal dynamics of founding *V. cholerae* populations along the intestine. (**a**) Corrected bottleneck population size (N_b', black dots) and bacterial load (cfu, red squares) at different loci during early, middle and late phases (~2 h; ~7 h; ~20 h post-infection) of infection from 19 animals from 12 independent litters. Open symbols represent N_b' values above the resolution limit or no detected colonies; dotted lines indicate the resolution limit for N_b' estimation. Sample medians are represented by horizontal lines. Corresponding N_b' and bacterial load from the same animal are aligned vertically and in the same sequential order. Significance was tested with one-sided Wilcoxon signed rank tests; * ($p < 0.05$) and

** ($p < 0.01$). (**b**) Genetic distance of populations during different phases of the disease. The genetic distance between I1 and I3 in the middle phase of the disease is significantly different from the early ($p = 0.026$) and the late phase ($p = 0.015$, both two-sided Wilcoxon rank sum test). (**c**) Model of the spatio-temporal dynamics of *V. cholerae* infection in the infant rabbit host. The bacterial burden is represented in the heat map (red: high; yellow and white: low); arrows indicate the direction of migration.